

Technical Bottlenecks and Process Optimization Pathways for AI Chip Wafer Manufacturing under Advanced Semiconductor Nodes

Qixu Pan

College of Engineering and Physical Sciences, University of Birmingham, Birmingham, UK

Abstract: With the rapid growth of AI workloads, there is now a high demand for high-end node wafer manufacturing, and at the same time, better power efficiency and higher yield are required. Identify the main technical bottlenecks in AI chip wafer manufacturing at advanced semiconductor nodes and put forward feasible process optimisation paths based on manufacturing integration. AI chips are not only the result of new-generation architecture but also depend on factors such as lithography technology, device structure, defect management, interconnect reliability, process fluctuation and yield optimisation to some extent in their actual performance and economic feasibility. Based on the theory of classical scaling, research in semiconductor manufacturing and recent reports on advanced-node processes have shown that the main problem is how to align performance scaling with manufacturability. Optimisation strategies based on design-technology co-optimisation, critical-layer process control, yield-oriented defect management and adaptive fab collaboration are also proposed. Therefore, the construction of advanced-node AI chip manufacturing will require the parallel development of device structure, process integration and manufacturing governance in the semiconductor industry.

Keywords: Semiconductor industry, AI chip, wafer manufacturing, advanced semiconductor nodes, process optimization, yield management.

1. Introduction

With the development of generative artificial intelligence and large-scale model training and high-throughput inference, the strategic importance of AI chips has been rising. AI accelerators in the current semiconductor industry are no longer marginal products for a few technical applications. They are now the foundations for cloud computing, edge intelligence, autonomous driving and industrial digitalisation. The competition of these programmes depends on their architecture and software ecosystems, as well as whether leading foundries can manufacture them at advanced semiconductor nodes with good yield, stable performance and reasonable cost.

As wafer manufacturing has advanced to an extent, so too has the problem. Transistor scaling in recent years is no longer linearly increasing the gain. Device physics, lithography accuracy, interconnect resistance, pattern fidelity, process variation, defect sensitivity and cycle-time pressure are all in a state of close connection. AI chips increase these tensions because they generally have large die sizes, dense compute arrays, high memory bandwidth requirements and strict power-performance goals. As a result, the problem is no longer how to create smaller transistors, but how to organise a high-integration manufacturing system with a narrow process window.

Therefore, the discussion of AI chips should not be limited to the level of algorithmic demand or computer architecture. Manufacturing capability is now a core bottleneck that determines whether advanced AI hardware can be realised. The same design goal that aims for a high-throughput may also increase overlay error, sensitivity to line-edge roughness, local variation, routing congestion and yield loss. The success of a high-end AI chip in practice depends on whether the wafer fabrication facilities can produce complex chips

without reducing production efficiency.

This paper investigates the technical bottlenecks and process optimisation paths for AI chip wafer manufacturing at advanced semiconductor nodes. Therefore, the problem of advanced-node manufacturing for AI applications should be viewed as a problem of coordinated process integration rather than one of individual process excellence. Section 2 introduces the technology basis for the advanced-node AI wafer manufacturing. Section 3: Analysis of the primary bottlenecks in fabrication. Section 4 proposes practical optimisation paths for the process and manufacturing organisation. Finally, present an overview of the semiconductor industry in the conclusion.

2. Technological Logic of Advanced-node AI Wafer Manufacturing

The technological foundation for advanced-node manufacturing started with the historical logic of scaling. Moore's first idea was to expect that integrated circuits would keep increasing in density over time, so the performance would rise and the cost drop for the whole semiconductor industry [1]. Dennard and his colleagues later formalised the scaling logic for MOSFET design, and showed why dimensional reduction could improve switching performance and integration efficiency under certain assumptions [2]. The two foundations have promoted the development of modern wafer production and are responsible for building a process-technology-based competitive advantage in semiconductor manufacturing.

Yet, advanced-node manufacturing is not just an extension of the previous scaling. As the complexity of the process increased, so too did the demand for yield management; thus, Stapper and Rosner have pointed out that high-quality integrated circuit production now requires systematic yield analysis, defect control and process feedback, not just

excellent design work [3]. Given the large die area and dense logic structure of AI chips, manufacturing losses are relatively more prominent in these compared to other traditional products. As the complexity of the node increases, so too does the technical requirement for a high-yield process and the economic cost of defects.

The basic process logic is also restricted by semiconductor device physics. Sze and Ng believe that the behaviour of devices at the nanoscale is governed by electrostatics, leakage control, channel engineering, material interactions and process-induced variations [4]. As AI chip designers add more transistors and reduce power consumption, fabrication needs to tackle the new physical constraints with higher-end process integration. Therefore, further reductions in the size of the features are no longer sufficient; we need to maintain the controlled behaviour of devices with billions of transistors under real-world fab conditions.

Bohr's review of Dennard scaling shows that the industry had reached a point where the classical scaling assumption was no longer valid automatically; therefore, more new materials, strained channels, architectural changes and complex manufacturing solutions had to be employed [5]. Therefore, the concept of progress in practice has also changed. The problem is no longer how to draw a smaller pattern; rather, it is how to maintain performance and power gain as variability, parasitics and integration complexity increase at a faster rate than shown in the nominal node label.

Lithography has the same problem. Levinson proposes that the new patterning process has a very narrow window, and resolution, depth of focus, stochastic effects, mask fidelity and overlay control are becoming more restrictive [6]. The lithographic constraints for AI chip wafer manufacturing are particularly severe, and the compute array and high-density interconnect fabric have little tolerance for critical-dimension drift or pattern irregularity. At the advanced node, lithography is no longer a production process but a structure constraint that limits the output of all scaled devices.

The demand for AI chips is also altering the technology of manufacturing. Jouppi and his co-authors have shown that the performance improvements of domain-specific AI accelerators are achieved by optimising dataflow, matrix operations and throughput-oriented computation [7]. However, these advantages also bring manufacturing pressure: large on-chip compute resources, dense routing and performance-per-watt goals all require cutting-edge process platforms. In light of the above, AI chips will not only use the high-end nodes but also actively increase the manufacturing demand for these nodes.

Advanced-node logic technology reports have identified the reasons for this integration problem. Auth and his colleagues have described a 10nm CMOS platform that requires FinFET optimisation, self-aligned quad patterning, contact-over-active-gate integration, and cobalt local interconnects, and have shown how multiple process innovations need to be coordinated in a single manufacturing architecture [8]. Loubet and his co-authors have also shown that stacked nanosheet gate-all-around transistors are a path beyond FinFET scaling, but such device innovation also increases the sensitivity of the process [9]. Therefore, the technology path for advanced-node AI manufacturing is a hybrid one that combines the goal of scaling with an increasing demand for structural manufacturability.

3. Technical Bottlenecks in AI Chip Wafer Manufacturing

The first is patterning complexity at the critical layer. Advanced nodes require very fine lithographic control, and as multi-patterning-intensive flows move towards higher demands for pattern fidelity, the stochastic and overlay risks are also increasing [6]. AI chips are generally large-die chips that contain many repetitive computation units; therefore, a local lithography error can result in a systematic yield loss rather than an isolated defect. Therefore, the bottleneck in wafer production is not whether a particular layer can be printed in principle, but rather whether it can be repeatedly and uniformly printed at a cost-effective rate across production batches.

The second is the integration of device architecture. Auth's 10nm logic platform shows that advanced-node fabrication requires the joint control of transistor architecture, patterning sequence, local interconnect metallurgy and contact integration [8]. Loubet's study on nanosheet gate-all-around structures has also shown that scaling beyond FinFETs imposes higher requirements for the formation of channel stacks, etch selectivity and sheet uniformity, and threshold control [9]. AI chips need to be high-frequency and power-efficient at the same time; therefore, any instability in device integration will affect binning yield, leakage control and thermal design margins.

A third bottleneck is fluctuation. Bohr and Young believe that CMOS scaling beyond the traditional path will be limited by variability, parasitic effects and power constraints; thus, a simple reduction in size will not lead to continuous performance improvement [10]. AI accelerators are more susceptible to the system-level impact of local variations due to their large-scale parallel arrays. Although the shift in threshold voltage, line resistance or contact integrity may be small at the single-device level, it can lead to a higher static power consumption, timing guardband inflation and reliability issues in the final product.

The fourth bottleneck is defect density and yield learning. Stapper and Rosner believe that yield improvement is not achieved by chance in the later stage, but by organising defect classification, space analysis and manufacturing control systematically [3]. Advanced AI chips are more significantly affected by slow yield learning because of large die size and high wafer cost. When defect mechanisms are associated with many critical layers, the fab should not rely on general improvement campaigns but needs product-specific yield diagnosis and quick cross-functional response.

A fifth bottleneck is the mismatch between AI design ambition and manufacturable process windows. AI chips require high-bandwidth, dense-computation and energy-efficient chips, but the above objectives often result in routing congestion, large reticle utilisation, tight electromigration exposure, and more severe thermal density [7]. Chips are logically feasible but physically fragile in terms of manufacturing. The discrepancy between the ideal architecture and the production-tolerance process is a relatively serious but often overlooked bottleneck in the advanced-node production stage.

Finally, the roadmap evidence shows that these bottlenecks will not be solved by simple expansion. The 2023 edition of IRDS shows that future progress in Moore's Law technology will face increasingly difficult interconnects, patterning challenges, variability control and integration costs, and will

not follow the previous scaling path [11]. AI chip wafer manufacturing is thus subject to a structural bottleneck problem. It is not a temporary defect in a single fab step, but the result of a series of narrow margins in device design, process integration, equipment capabilities, yield control and cost governance.

4. Process Optimization Pathways

The first optimisation path is to enhance design-technology co-optimisation. Given that the manufacturing difficulty of AI chips is strongly determined by large die size, dense compute fabric and aggressive power targets, process optimisation cannot start after the design is finalised. The logic of scaling from Moore and Dennard already shows that the progress of devices depends on the alignment of design assumptions with physical implementation [1-2]. Advanced-node AI products should integrate early manufacturing feedback on pattern density, routing hotspots, power-grid stress and critical-area sensitivity in practice. Thus, the risk of an architecturally superb Design being impractical in construction is reduced.

The second is to use essential-process control rather than widespread optimisation. According to lithography theory, the success of advanced patterning is highly sensitive to narrow windows in overlay, focus, stochastic control and resist behaviour [6]. Therefore, given the above, Fabs should focus their metrology, statistical process control and excursion prevention efforts on the layers that account for a relatively large proportion of the final yield, rather than focusing solely on the easy-to-monitor ones. For AI wafers, this usually refers to a closer integration of lithography, etch and inspection strategies around transistor fins or sheets, contact definition and the most congestion-sensitive interconnect layers.

The third way is device-aware process integration. Reports on 10 nm FinFET technology and stacked nanosheet gate-all-around structures show that the success of advanced nodes requires process modules to be optimised as interdependent systems rather than isolated steps [8, 9]. Etch selectivity, spacer quality, contact resistance, metallization selection and thermal budget all need to be controlled in relation to the final transistor and interconnect structure. For the manufacturing of AI chips, such integration should be associated with product-specific power-density requirements and target operating windows, as the acceptable variability for one logic product may not meet the standards of another.

The fourth way is yield-oriented defect governance. Yield management literature indicates that high-value products require the establishment of an orderly learning loop connecting inspection, failure analysis, spatial signature recognition and process correction [3]. Therefore, the advanced AI chips should be viewed as high-speed learning vehicles for defect discovery rather than as regular volume products awaiting generalisation. A joint dashboard for foundries and product teams needs to be introduced to display defect Pareto analysis, wafer map correlation and line-of-sight ownership of recurring loss mechanisms. Shorten the cycle of fab event, diagnosis and correction.

The fifth way is variability reduction by adaptive margining and process-window engineering. Device physics and post-Dennard scaling analysis have both shown that variability can no longer be handled by downstream binning or larger guardbands alone [4, 5, 10]. Optimisation of the process should aim to reduce variability at its source by improving pattern fidelity, strengthening stable material

integration, enhancing statistical control over critical dimensions, and adding feedback from in-line metrology and electrical testing. AI chips need to be energy-efficient and meet frequency goals, so adding a lot of design conservatism is not ideal.

The sixth way is manufacturing strategy at the ecosystem level. AI chip production falls under the general category of the semiconductor industry, which is highly capital-intensive, concentrated in equipment, and changes rapidly. According to the IRDS roadmap, further development of the nodes will require joint efforts by all devices, interconnects, metrology and the whole system [11]. Optimisation cannot be achieved by fab-floor methods alone, therefore. Foundries, EDA companies, equipment manufacturers, IP teams and top AI chip customers will also need to coordinate the process assumption, design rule and yield expectation of the integrated system through cooperation.

Finally, redefine the measure of success for advanced-node AI wafer manufacturing. Jouppi's analysis of domain-specific AI acceleration shows that performance per watt and application efficiency are necessary for the value of AI hardware [7]. Therefore, the manufacturing optimisation should judge wafers based on the stability of usable die output, predictable power behaviour and scalable production economics, rather than just transistor density. At an advanced semiconductor node, the leading AI chip is one that can turn technical innovation into mass-production-ready industrial products via a reliable process.

Add process-data analysis to the regular optimisation work for AI chip manufacturing. Inline metrology, electrical test results, wafer maps, equipment logs and defect images can be combined for weak process window and recurrent failure signature identification. The above data-driven feedback is not to replace physical knowledge but rather to help engineers quickly identify the reasons for the change and modify the process recipe before a small deviation results in yield loss. The AI chip with an advanced node is extremely sensitive to small local variations and may thus suffer from timing errors, leakage, power delivery problems, and functional binning across a large die. Link manufacturing data with product-level performance indicators to enable more predictive optimisation of the fab, reduce the yield-learning cycle, and support stable high-volume production.

5. Conclusion

Cutting-edge-node AI chip wafer manufacturing is now required for scaling semiconductors, innovating devices, and boosting industrial strength. The problem is no longer the individual pursuit of smaller features, but rather the coordinated control of lithography accuracy, equipment integration, variation, defect learning and manufacturing costs in a period of tightening process windows. AI chips increase the demand for density, power efficiency and yield stability due to the high-end architecture of these chips.

The principal reasons for the bottlenecks in the structure of AI chip wafer production are not random. Patterning complexity, architecture integration, variability, defect sensitivity, and the widening gap between design ambition and manufacturable process window all limit the effectiveness of advanced semiconductor nodes. Therefore, the optimisation of the process should combine design-technology co-optimisation, critical-layer control, yield-centred governance, variability reduction and ecosystem-level collaboration.

In short, the next round of competition in the semiconductor industry will not be about announcing a higher-end node, but rather which company can reliably scale its advanced-node technology to high-volume manufacturing for AI chips. Therefore, optimisation of the process is not a relatively small-scale engineering adjustment. It is the practical path by which advanced semiconductor technology becomes stable industrial capacity.

References

- [1] G. E. Moore. (1965). Cramming more components onto integrated circuits. *Electronics*, 38(8), 114–117.
- [2] Dennard, R. H., Gaensslen, F. H., Yu, H.-N., Rideout, V. L., Bassous, E., & LeBlanc, A. R. (1974). Design of ion-implanted MOSFET's with very small physical dimensions. *IEEE Journal of Solid-State Circuits*, 9(5), 256–268. <https://doi.org/10.1109/JSSC.1974.1050511>
- [3] Stapper, C. H., & Rosner, R. J. (1995). Integrated circuit yield management and yield analysis: Development and implementation. *IEEE Transactions on Semiconductor Manufacturing*, 8(2), 95–102. <https://doi.org/10.1109/66.388016>
- [4] Sze, S. M., & Ng, K. K. (2006). *Physics of semiconductor devices* (3rd ed.). John Wiley & Sons. <https://doi.org/10.1002/0470068329>
- [5] Bohr, M. (2007). A 30 year retrospective on Dennard's MOSFET scaling paper. *IEEE Solid-State Circuits Society Newsletter*, 12(1), 11–13.
- [6] Levinson, H. J. (2011). *Principles of lithography* (3rd ed.). SPIE Press. <https://doi.org/10.1117/3.865363>
- [7] Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., et al. (2017). In-datacenter performance analysis of a Tensor Processing Unit. *Proceedings of the 44th Annual International Symposium on Computer Architecture*, 1–12. <https://doi.org/10.1145/3079856.3080246>
- [8] Auth, C., Aliyarukunju, A., Asoro, M., Bergstrom, D., Bish, D., Bost, M., et al. (2017). A 10nm high performance and low-power CMOS technology featuring 3rd generation FinFET transistors, self-aligned quad patterning, contact over active gate and cobalt local interconnects. *2017 IEEE International Electron Devices Meeting*, 29.1.1–29.1.4. <https://doi.org/10.1109/IEDM.2017.8268472>
- [9] Loubet, N., Hook, T., Montanini, P., Yeung, C.-W., Kanakasabapathy, S., Guillorn, M., et al. (2017). Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET. *2017 Symposium on VLSI Technology*, T230–T231. <https://doi.org/10.23919/VLSIT.2017.7998183>
- [10] Bohr, M., & Young, I. A. (2017). CMOS scaling trends and beyond. *IEEE Micro*, 37(6), 20–29. <https://doi.org/10.1109/MM.2017.4241347>
- [11] IEEE International Roadmap for Devices and Systems. (2023). *International roadmap for devices and systems: 2023 Edition*. IEEE.